立项依据：

包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录(字数不少于：800)

## 研究目的及意义

在过去几十年中，互联网极大地促进了电子商务的发展，同时互联网也成为地下市场的主要平台，网络犯罪分子在这里交换用于犯罪的产品和服务。2013年，全球网络犯罪造成了3750亿美元的损失，几乎等同于全球毒品交易的数量。

地下论坛是网络罪犯的通信枢纽，帮助他们推广攻击工具包和服务，协调他们的行动，交换信息并寻求合作。例如Silk Road论坛，拥有3万到5万活跃用户，是毒品和其他非法毒品交易的滋生地，每天有两到三百份通讯记录。这些记录提供了对网络犯罪方式、犯罪分子战略、能力、基础设施和商业模式的深刻洞察，甚至可以用来预测他们的下一步行动。

不仅在地下论坛，一些广为人知的社交平台也被不法分子用于交易通讯。近年来，大型社交媒体公司一直在雇佣内容版主，对社交平台上的内容进行审查。伴随着互联网的发展，许多社交平台每天都产生了海量的通讯记录，完全依靠人工处理的方式难以应对如此大的数据量。目前的许多平台都开始使用一些自动化的方法来帮助审查，使用最多的方法就是创建一个“禁用词列表”，通过列表来对通讯内容进行过滤。

由于地下市场的排他性，网络罪犯发展了一套独特的语言系统，违法者会使用一些黑暗术语来对外隐藏具体交易内容或规避社交平台的内容审查。黑暗术语用来指定产品、服务和其他网络犯罪特定概念。这种术语往往是一些看起来很普通、看很天真、却有秘密含义的词语。罪犯们将其用来隐藏正在讨论的内容。例如，毒贩经常用“ice”代替Methamphetamine（冰毒），“pants”代替Herion（海洛因）。此类欺骗性内容使得违法者之间的通信不容易被自动化系统和审查人员发现。因此，自动发现和理解这些黑暗术语对于理解各种网络犯罪活动和减轻它们所构成的威胁非常有价值。

## 国内外研究现状和发展趋势

在国外文献中，黑暗术语检测的具体任务已经在许多框架下进行了研究。

G. Durrett等人[6]、[7]通过自动、自上而下的方法分析地下论坛生成高级信息。该方法将信息提取和命名实体识别相结合并具体提出了一种数据标注方法，并利用标记数据训练基于监督学习的分类器。然而结果在很大程度上取决于注释的质量，正如几位研究人员所示，该模型在跨域数据集中的表现并不好。

K. Yuan等人提出了一种名为CantReader的系统，旨在自动识别网络犯罪市场中的“黑色术语”。CantReader采用基于神经网络的嵌入技术来分析单词的语义，并检测背景语料库（如维基百科）中的上下文与目标语料库中的上下文显著不同来确认黑暗术语的候选词。因此，它需要一个“黑暗”语料库（如丝绸之路匿名在线市场论坛）、一个混合语料库（如Reddit）和一个良性语料库（如英语维基百科）作为输入。可能是由于词多义的影响，系统在其他数据集上的表现并不好。

C. Felt 等人[17]提出通过情绪分析来检测委婉语。它使用语义词典归纳的自举算法识别一组委婉语候选词。虽然这种方法似乎合理且可行但它需要额外的手动筛选过程来细化候选对象，因此无法满足所期望的自动、大规模检测的要求。

Magu等人[11]和Taylor等人[40]提出了两种利用单词嵌入和社区检测算法的算法。Magu等人[11]通过特征向量中心度的排名度量生成了一组术语。由于算法的本身性质，这种方法需要一个开始的术语种子来帮助寻找其他的术语。Taylor等人[40]创建了神经嵌入模型，捕捉单词的相似性，使用图扩展和PageRank分数引导初始种子词，并最终丰富引导词，从表现类似委婉语的字典术语中学习。Magu和Taylor的方法在跨领域数据上表现不佳，Talor方法使用了大量的筛查方法，却没有做消融实验，无法确定各个方法的具体作用。

Wanzheng Zhu等人提出了一种自监督方法来实现对委婉语的检测，通过术语与指代的禁用词的相似性，结合当前NLP领域流行的MLM掩码模型，计算术语与禁用词上下文相似的可能性，从而确定某个禁用词的一组候选语。通过同样的方法结合两个分类器来确定术语的语义。该方法思路新颖，但精确度一般且在检测时由于无监督存在的标签缺失的可能没有设计合理的召回率指标。

国内也有一些针对中文语境的术语检测和研究。

Yang等人[39]构建了一个关键词检测和扩展系统（KDES），并将其应用于中国最大的搜索引擎百度的搜索结果上。KDES旨在通过检查相关的搜索结果来推断是否应该阻止搜索关键字。这种方法需要具有能够访问远程域信息（即百度搜索引擎），不适合我们所需要分析的社交平台内容和地下论坛内容。

Kangzhi Zhao等人也从术语与禁用词上下文相似角度出发，通过word2vec和LDA等无监督方法衡量术语与禁用词的相似性确定术语。数据来源为在通讯软件qq中存在的黑色交易群聊中收集聊天记录作为数据。术语与禁用词的检测中存在多义问题和语料分布不均匀问题都会对word2vec和LDA方法带来较大的影响，不符合我们的期望。

## 主要参考文献目录

G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, R. Portnoff, S. Afroz,D. McCoy, K. Levchenko, and V. Paxson, “Identifying products in onlinecybercrime marketplaces: A dataset for fine-grained domain adaptation,”inProceedings of Empirical Methods in Natural Language Processing(EMNLP), 2017, pp. 2598–2607

R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, “Tools forautomated analysis of cybercriminal markets,” inProceedings ofInternational Conference on World Wide Web (WWW), 2017, pp. 657–666.

K. Yuan, H. Lu, X. Liao, and X. Wang, “Reading thieves’ cant: auto-matically identifying and understanding dark jargons from cybercrimemarketplaces,” inProceedings of 27th USENIX Security Symposium,2018, pp. 1027–1041.

R. Magu and J. Luo, “Determining code words in euphemistic hatespeech using word embedding networks,” inProceedings of the 2ndWorkshop on Abusive Language Online (ALW2), 2018, pp. 93–100.

J. Taylor, M. Peignon, and Y .-S. Chen, “Surfacing contextual hate speech words within social media,”arXiv preprint arXiv:1711.10093, 2017.

Zhu W, Gong H, Bansal R, et al. Self-supervised euphemism detection and identification for content moderation[J]. arXiv preprint arXiv:2103.16808, 2021.

H. Yang, X. Ma, K. Du, Z. Li, H. Duan, X. Su, G. Liu, Z. Geng,and J. Wu, “How to learn klingon without a dictionary: Detection andmeasurement of black keywords used by the underground economy,”inIEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp.751–769.

Zhao K, Zhang Y, Xing C, et al. Chinese underground market jargon analysis based on unsupervised learning[C]//2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016: 97-102.

# 研究思路

研究内容、研究目标、关键问题：说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容(字数不少于：2500)

近年来随着网络犯罪的兴起，大量社交媒体和网络社区开始花费更多时间和人力到内容审查上。内容审查主要指的是平台人员确定平台上出现的内容是否适度(主要是出现在公共社区的帖子以及搜索引擎结果等)，以及是否应当保留或删除。不适当内容包含许多方面，如仇恨言论、引战、色情信息等。但对于网络审查人员来说，最重要的是那些关于网络犯罪交易的内容，如毒品交易、武器交易、性交易等。庞大的网民数量和交流需求导致了海量的待审查数据。庞大且不乏恶意的帖子很容易冲垮审查人员的身体和精神。审查自动化是目前许多平台的一项重大需求。自动化审查的最简单的一种方式是基于“禁用词列表“的禁用词系统，通过检索禁用词来完成内容审查。但是这种方式很容易被规避，违法者可以通过发明新的词来代替禁用词（我们称之为黑暗术语），而这些被用来替代的黑暗术语本身往往是一些具有其他无害含义的词，不能被无条件过滤。我们需要联系词的上下文来进行判断，例如在中文中会使用叶子来代替”大麻“，使用”飞行“代替”吸毒“。为了规避审查，几乎所有的黑暗术语都有无害的含义。”美国进口高纯度飞友必备叶子限量出货不懂的请自行问问哥，骗图的别来“，这是来自某社交平台的毒品交易聊天记录，显然即使我们不明白”叶子”的具体含义，我们仍然能确定这是一个黑暗术语。这表明即使不确定具体含义，仍然能够根据上下文确定黑暗术语。

为了帮助现有的禁用词系统来检测这类规避情况，我们需要建模上下文信息来进行黑暗术语识别。基于深度学习的NLP技术在建模上下文信息上的作用已被广泛认可，也是本课题拟采用的技术。我们的目标是配合禁用词系统来达到更好的审查效果，基于此将本课题任务分为两类1.发现黑暗术语 2.判断黑暗术语含义。任务一的目的是构建一个模型能够根据内容来发现可能的黑暗术语。任务二的目的是确定黑暗术语在具体句子中的语义。任务一的主要挑战是区分术语的行话含义和无害。例如”美国进口高纯度飞友必备叶子限量出货不懂的请自行问问哥，骗图的别来“中，“叶子”可以指树叶。但我们可以根据上下文判断这是不可能的，因为没有人会关心树叶是否是纯的。过去的研究依赖于静态单词嵌入（例如，word2vec[13]），希望通过比较在不同语料上（如黑色交易语料和普通交流语料）上的差异来检测黑暗术语，这种基于语料级别的比较容易受到语料中的词的多义表达干扰。一类较新的上下文自动嵌入（例如，BERT]）为单词出现的每个上下文学习不同的单词表示，因此它们不会混淆同一单词的不同含义。任务二的目的是当禁用词系统根据任务一发现的黑暗术语得到大量句子时，对句子中黑暗术语的具体语义进行判断，筛查假阳性的句子（即用了无害含义的句子）。任务二的研究相对较少，大部分研究都没有具体去对黑暗属于含义进行鉴别。Yuan等人通过识别委婉语的上位词（例如，它是指药物还是人）来解决一个相关问题。Wanzheng Zhu通过使用两个分类器来对黑暗术语计算含义的概率分布，但准确的不足以应付实际的排除假阳性的需求。从更一般的意义上讲，委婉语识别的任务也与未知词的词义发和词义消歧有关。词义发现的目的是通过生成一个定义句来理解一个未知单词的意义，而词义消歧则侧重于在给定一组候选词义的情况下，识别一个句子中使用了单词的哪个词义，并严重依赖语言学家和词典编纂者创建的带有词义标记的参考语料库。但是，这两种方法都无法捕捉同一类别中语义相似的一组目标关键字之间的细微差异。

随着中国网络的不断发展，同样的情况和研究在出现。许多网络安全研究人员已经对术语识别进行了研究。然而，现有的研究大多集中在英语专业术语上，而没有基于汉语专业术语的研究。本课题希望能在中文语境下对黑暗术语识别进行探索。由于中英文之间的差距，目前的中文行话检测还存在许多问题。比如由于语言特征，基于英文的许多特征并不能用于中文；英文中的黑暗术语往往是单词形式出现（如ice代表海洛因），而在中文中的黑暗术语往往以短语形式，单词级别的检测和短语级别的检测使许多基于英文的研究结果无法套用到中文中去；违法者们使用的语言表达形式往往较为多样，要捕捉这样丰富的上下文信息需要大量的标注数据，会带来较大的标注负担。同时新的表达形式在不断出现，也意味着新的标注任务，循环无休。

自监督是一种减轻标注负担的好的方法。自监督是一种无监督学习形式，数据本身提供监督。自我监督的目的是利用大量未标记的数据（如自由文本、图像），从数据本身构建一个有监督的学习任务来预测数据的某些属性。例如，训练一个文本预测模型，可以获取一个文本语料库，屏蔽句子的一部分，然后训练模型预测被屏蔽的部分；此工作流从未标记的数据创建有监督的学习任务。自监督已广泛应用于语言建模表征学习、机器人学、计算机视觉和强化学习。Wanzheng Zhu等人提出了一种自监督方法来实现对黑暗术语的检测，通过术语与指代的禁用词上下文的相似性，结合当前NLP领域流行的MLM掩码模型，计算术语与禁用词上下文相似的可能性，反复使用MLM模型来确定某个禁用词可能的黑暗术语。同时使用一个粗分类器排除一些假阳性结果，再结合细分类器得到一个黑暗属于含义的概率分布。但该研究使用的MLM模型并不能直接用于短语的检测，同时模型本身的检测精度并不高，仍然由许多地方可以完善。该模型使用了一些地下论坛的语料数据来对mlm进行微调，以及带有禁用词的句子来得到候选黑暗术语。但是由于网络审查的逐渐普遍化，两种类型的语料都变得难以获取（作者2021年的论文，数据却来自2018，缘于许多论坛的被监察消失）。首先是地下论坛的语料数据，我们可以称其为黑暗语料，里面可能包含着大量的违法者特有的语言风格以及未被识别到的黑暗术语。目前的地下论坛基本以英文交流为主，如丝绸之路、xxx等，在中文环境下也很难在社交平台上找到这样一个能产生大量黑暗语料的群体（某些平台存在非法交易群聊，但这并不适用于所有平台，而且群聊随时有可能被监管到，大部分平台能稳定捕捉到的数据都来自或大或小的公共社区的分享）。同时对于带禁用词的句子，这部分数据由于中文环境的审查的普及，几乎无法获取。对于一个深度学习模型而言，数据是否足够很大程度上会影响模型的鲁棒性，而数据是否够新则决定模型能否捕捉语料中上下文的新的变化。我们需要考虑如何对这两类语料进行补充，使模型表现更加出色同时能适应新的变化。

研究方案：

包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性(字数不少于：800)

课题特色

本研究课题的特色与新颖(创新)之处(字数不少于：500)